AI-Generated Text Detection

CS5100 Final Project

Student Name(s): [Your Name]

Date: [Submission Date]

# 1. Introduction

This project focuses on detecting whether a given paragraph of text was written by a human or generated by an AI model (e.g., GPT). The rapid advancement of large language models introduces new challenges, such as misinformation and academic dishonesty, making AI-generated text detection an important area of research.

# 2. Background

Existing research on AI text detection includes approaches such as OpenAI's AI Text Classifier, GPTZero, and DetectGPT. These tools use machine learning and linguistic patterns to differentiate AI-generated content from human-written text. In this project, we explore both classical machine learning approaches (TF-IDF + Logistic Regression) and modern transformer-based architectures (BERT).

# 3. Methodology

The HC3 dataset, which contains human and AI-generated answers to the same questions, was used. We preprocessed the dataset by flattening human and AI responses, assigning binary labels (0 = human, 1 = AI), and saving the processed data for training.  
  
Two approaches were tested:  
- \*\*Baseline Model:\*\* TF-IDF vectorization combined with Logistic Regression.  
- \*\*Transformer Model:\*\* Fine-tuning BERT for binary text classification.  
  
Tools and libraries used include HuggingFace Transformers, scikit-learn, PyTorch, pandas, and matplotlib.

# 4. Results

The baseline TF-IDF + Logistic Regression model achieved [Insert Accuracy/F1-Score]. The fine-tuned BERT model achieved [Insert Accuracy/F1-Score], showing a significant improvement over the baseline.  
  
Below is an example confusion matrix and classification report from the baseline model:  
  
[Insert confusion matrix image and classification report here]

# 5. Discussion

While BERT provided higher accuracy, it required significantly more computational resources and training time. The TF-IDF model, although simpler, performed well for shorter text. One limitation is that AI-generated text continues to evolve, which may reduce detection accuracy over time. Future work could include training on a wider range of AI model outputs or using ensemble methods.

# 6. Conclusion

This project demonstrates that transformer-based approaches like BERT outperform classical models for AI text detection. We learned the importance of dataset preprocessing, model fine-tuning, and evaluation metrics. The work can be extended to real-time detection tools and web-based applications.

# 7. References

[1] HC3 Dataset - https://huggingface.co/datasets/Hello-SimpleAI/HC3  
[2] HuggingFace Transformers - https://huggingface.co/transformers/  
[3] scikit-learn - https://scikit-learn.org/  
[4] DetectGPT - https://arxiv.org/abs/2301.11305